9. Testing

From **"**Verification, Validation, and Evaluation of Expert Systems, Volume I"

This chapter discusses how a simple experiment can be designed to test whether an expert system satisfies a specification.

Simple Experiments for the Rate of Success

The most common statistic measuring how well a system satisfies a specification is to observe the expected fraction of inputs on which the system will satisfy the specification. One can estimate this fraction of an experiment based on the following steps:

1. Select a data sample.
2. Run the expert system on the data sample.
3. Analyze the experimental data.

Selecting a Data Sample

Each specification for the expert system is of the form:

> If the input satisfies certain conditions, then the output satisfies certain other conditions.

A sample of N data items for a specification is a set of N data items that satisfies the conditions in the if part of the specification. Furthermore, the sample should satisfy the following additional condition:

> If x is a variable which is thought to affect the reliability of the expert system on the specification, the distribution of x in the sample should approximate the distribution of x in the underlying population.

There are several ways to collect a sample:

random subsample: If a sample of data was put aside for testing during the initial phase of the system lifecycle, the experimenter can draw a random subsample from this sample.

monitoring: Potential inputs can be collected from the environment where the expert system is to perform. A subset of the observed inputs that satisfy the conditions of the specification to be tested becomes the sample for the experiment.

generated input data: Where actual data is not available or practical, a computer program can be used to randomly generate data satisfying the input conditions of the specification.

The size of the sample that should be selected is estimated below.

If a specification has been proved to be satisfied, the existence of the proof may increase the reliability achieved by a test. This effect is also discussed below.

Estimating a Proportion (Fraction) of a Population

If the expert system is run on N data items, and it satisfies the specification on K of those items, then:

K/N = the experimental point (i.e., single number) estimate of the proportion of the underlying population satisfying the specification

If the sample size N is sufficiently large, the distribution of sample proportions (the values of K/N) is approximately normal. This occurs when both the following conditions are true:

$$K = N*(K/N) > 5 \qquad (9.1)$$

$$N*(1-K/N) > 5$$

When this is true, the standard error of the proportion is

$$s\_e(K/N) = \text{sqrt}((K/N)*(1-K/N)/N) \qquad (9.2)$$

When the conditions (9.1) for normality are not satisfied, the Poisson distribution, discussed below can be used to estimate the satisfaction of a specification.

The Confidence Interval of a Proportion

In this section the goal is to find an interval of proportions (fractions) of a population since most of the time the observed satisfaction of the specification for a new sample will be in the interval. In particular, the goal is to find an interval such that the probability of the observed satisfaction being in the interval is (sat) for (sat) close to 1.

The steps in computing the interval are:

1. Conduct the experiment to test the specification. Observe:
   - The sample size N.
   - The number of times K the specification was satisfied on the sample.
   - Conduct enough trials so that the requirements for approximate normality are satisfied.
2. Compute s_e(K/N)
3. From a statistical table, find the standard normal deviate (snd) of sat, often called the "z-score" and denoted by z.

The standard normal deviate is a multiple of the standard error marking out a central region of the normal distribution that contains a given fraction of the total area (which is 1) under the normal distribution. In particular, z(sat) is the number such that the area under the normal distribution between -z(sat) and z(sat) is sat, i.e.,

$$\text{INTEGRAL}_{-z(sat)}^{z(sat)} \text{normal}(x)dx \qquad (9.3)$$

where normal(x) is the standard normal distribution,

$$n(x) = (1/2*pi)* \exp(-x^2/2) \qquad (9.4)$$

While there is no closed form for z, tables of z-scores are widely available in statistics texts; typical values are shown below:

sat	z(sat)
50%	0.68
75%	1.15
90%	1.65
95%	1.96
98%	2.33
99%	2.58
99.5%	2.81
99.8%	3.08

When K successes are observed in N trials, the sat confidence interval is

$$K/N \; +/- \; z(sat)*s\_e(K/N) \qquad (9.5)$$

Choosing Sample Size

The goal for a system developer is often to show that a system will perform at least as reliably as some threshold. Statistically, this means that with a confidence of at least C, a specification is satisfied in at least fraction F of a sample on which the specification applies. A typical statement of this form is:

> The expert system correctly diagnoses pavement maintenance remedies at least 90 percent of the time with 95 percent confidence.

This means that if another experiment using the same sample size was conducted, at least 95 percent of the time the measured fraction on which the specification is satisfied would be at least 90 percent.

Given a desired fraction F and a confidence level C, the user can obtain the size of sample needed to achieve these parameters in the following way:

1. Conduct a small initial experiment to estimate the fraction on which the specification is satisfied. This initial estimate will be denoted F0. If F0<F and the sample size of the initial experiment guarantees that there is reasonable confidence in F0, the expert system does not satisfy the proportion F. If F0 is equal or only slightly larger than F, the size of the experiment needed to narrow the confidence interval around F0 to exclude F will be unreasonably large; in practice, it will be impossible to statistically validate the satisfaction with proportion F and confidence C.

2. Given that F0 >F, compute:

$$s\_e = (F0-F)/z(C) \qquad (9.6)$$

To achieve F and C, choose a sample size such that the standard error is less than or equal to s_e. This means choosing an N such that:

$$\text{sqrt}(F0*(1-F0) / N) =< s\_e \qquad (9.7)$$

or

$$N >= F0*(1-F0)/s\_e\text{^}2 \qquad (9.8)$$

For example, if:

preliminary experimental proportion (F0) = 93%

minimum acceptable proportion (F) = 90%

confidence interval = 95%

then

$$s\_e = (93\%-90\%)/z(95\%) = 0.03/1.96 = 0.153$$

and

$$N >= 93\% * (1-93\%) / 0.153\text{^}2 = 277.9$$

This estimate of sample size is approximate, because the preliminary proportion F0 used in the computation is only the result of a small preliminary experiment, and will contain some random error. Therefore, the experimenter should, if possible, design an experiment so that an initial experiment can be continued by testing more data. This is possible provided that the probability of drawing any data item in the continuation of the experiment is the same as drawing that data item in the initial experiment.

Estimating Very Reliable Systems

For systems that do not fail often it is difficult in practice to observe the five or more failures that causes the proportion to be approximately normally distributed. In this case the Poisson distribution should be used as follows to estimate a confidence interval for the satisfaction proportion.

The Poisson distribution describes the number of occurrences of some random event in given interval of time or region of space. For example, the number of fish over any square meter of a lake, where the lake bottom is uniformly attractive to fish, is approximately Poisson distributed.

The formal requirements for an occurrence to be Poisson distributed include:

- Each occurrence is independent of the others.
- Each interval can potentially contain an infinite number of occurrences.

In practice, the second requirement can be approximated if a large number of occurrences can occur in a region; what is "large" for this purpose will be estimated below.

If the average number of occurrences in a region is L, the probability of finding k occurrences is:

$$P(k) = \exp(-L)*L^k/k! \qquad (9.9)$$

The probability of K or more occurrences is:

$$\text{SUM } \exp(-L)*L^k/k! \qquad (9.10)$$

$$k>=K$$

a series that converges geometrically once $L/k < 1$.

For testing a specification, a region is defined to consist of N trials, where N is a number such that N or more occurrences is very unlikely, as computed by (9.10).

The requirement that a specification is satisfied at a proportion at least F, means that

$$(N\text{-Fail})/N > F \qquad (9.11)$$

where N is the number of trials in a region, and Fail is the number of failures observed in N trials. This means that the number of failures Fail should satisfy is:

$$\text{Fail} < (1\text{-}F)* N \qquad (9.12)$$

This says that the number of failures should be less than the acceptable failure rate, 1-F, times the number of trials in a region. Using (9) the user can compute the probabilities of observing failure rates satisfying (9.12). Denote the sum of these probabilities by

$$P = \text{SUM } P(k) \qquad (9.13)$$

$$k < (1-F)*N$$

Then if P >= F, the expected success rate is at least F.

How a Proof Increases Reliability

Suppose that in a Hoffman region a specification has been proved and verified in a single experimental trial.

The question to be asked then is:

> What is the probability that the specification would fail on a new trial with inputs in the Hoffman region ?

By the definition of a Hoffman region, all atomic formulas that determine the computational path of the system have the same truth values for the inputs of the second trial. Therefore, the output on the new trial should be identical to that of the first on which the specification was satisfied. The only way for the outcome of the second trial to be different is for a system error to have occurred. Therefore, the probability of a failure on a Hoffman region for which both a proof and a single trial experimental verification is available is the probability of an underlying computer hardware or system software error occurring during the computation. As the Pentium bug illustrates this is a small, but non-zero probability.

In order for a fielded system to perform reliably, the probability of a computer system error must be kept small. However computer system error probability applies approximately equally to all knowledge bases. Therefore, once the underlying reliability of the computer system is established, resources should not be expended testing for this error. In particular, where a proof exists, one experimental trial per Hoffman region is sufficient to verify a specification with probability 1-Fc, where Fc is the probability of a computer system error.